

# A Probabilistic Model of Transcription Dynamics applied to Estrogen Signalling in Breast Cancer Cells

Ciira wa Maina<sup>1,\*</sup>, Filomena Matarese<sup>2</sup>, Korbinian Grote<sup>3</sup>, Hendrik G. Stunnenberg<sup>2</sup>, George Reid<sup>4</sup>, Antti Honkela<sup>5</sup>, Neil D. Lawrence<sup>1,\*</sup>, Magnus Rattray<sup>6,\*</sup>

**1 Department of Computer Science, University of Sheffield, Sheffield, UK**

**2 Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen, NL**

**3 Genomatix Software GmbH, Muenchen, Germany**

**4 Institute for Molecular Biology, Mainz, Germany**

**5 Helsinki Institute for Information Technology, Department of Computer Science, University of Helsinki, Helsinki, Finland**

**6 Faculty of Life Sciences, University of Manchester, Manchester, UK**

**\* E-mail: c.maina@sheffield.ac.uk, magnus.rattray@manchester.ac.uk, n.lawrence@sheffield.ac.uk**

## Abstract

Gene transcription mediated by RNA polymerase II (pol-II) is a key step in gene expression. The dynamics of pol-II moving along the transcribed region influences the rate and timing of gene expression. In this work we present a probabilistic model of transcription dynamics which is fitted to pol-II occupancy time course data measured using ChIP-Seq. The model can be used to estimate transcription speed and to infer the temporal pol-II activity profile at the gene promoter. Model parameters are determined using either maximum likelihood estimation or via Bayesian inference using Markov chain Monte Carlo sampling. The Bayesian approach provides confidence intervals for parameter estimates and allows the use of priors that capture domain knowledge, e.g. the expected range of transcription speeds, based on previous experiments. The model describes the movement of pol-II down the gene body and can be used to identify the time of induction for transcriptionally engaged genes. By clustering the inferred promoter activity time profiles, we are able to determine which genes respond quickly to stimuli and group genes that share activity profiles and may therefore be co-regulated. We apply our methodology to biological data obtained using ChIP-seq to measure pol-II occupancy genome-wide when MCF-7 human breast cancer cells are treated with estradiol (E2). The transcription speeds we obtain agree with those obtained previously for smaller numbers of genes with the advantage that our approach can be applied genome-wide. We validate the biological significance of the pol-II promoter activity clusters by investigating cluster-specific transcription factor binding patterns and determining canonical pathway enrichment.

## Author Summary

Cells express proteins in response to changes in their environment so as to maintain normal function. An initial step in the expression of proteins is transcription, which is dependent upon a number of sequential events, such as recruitment of pol-II to the transcriptional start site, activation of pol-II through phosphorylation of its C-terminal domain, elongation of the nascent transcript through the transcribed region and termination. To understand changes in transcription arising due to stimuli it is useful to model the dynamics of transcription. We present a probabilistic model of transcription dynamics that can be used to compute RNA transcription speed and infer the temporal pol-II activity at the gene promoter. The inferred promoter activity profile is used to determine genes that are responding in a coordinated manner to stimuli and are therefore potentially co-regulated. Our model can be parameterised using data from high-throughput sequencing assays, such as ChIP-Seq and GRO-Seq, and can therefore be applied genome-wide in an unbiased manner. We apply the method to pol-II ChIP-Seq time course data from breast cancer cells stimulated by estradiol in order to uncover the dynamics of early response genes in

this system.

## Introduction

Transcription mediated by RNA polymerase II (pol-II) is an essential process in the expression of protein-coding genes in eukaryotes. Transcription is dependent upon a number of sequential and dynamic events, such as recruitment of pol-II to the transcriptional start site, activation of pol-II through phosphorylation of its C-terminal domain, elongation of the nascent transcript through the transcribed region and termination [1]. Each of these steps may be rate-limiting and can therefore affect the level of gene expression. In this manuscript, we describe a simple probabilistic model of transcription whose parameters can be learned using time-series data such as pol-II ChIP-Seq data [2] or nascent transcript measurement by GRO-Seq that reports markers of transcriptional activity [3]. This model can be used to identify transcriptionally engaged genes, estimate their transcription rates and infer transcriptional activity adjacent to the promoter. The transcriptional dynamics of estrogen responsive genes in a breast cancer cell line were described by fitting this model to sequential pol-II ChIP-seq datasets.

Chromatin Immunoprecipitation, in conjunction with massively parallel sequencing (ChIP-seq) evaluates interactions between proteins and DNA, and, for example, can be used to monitor the presence of pol-II on DNA. Estimating the amount of pol-II associated with a transcribed gene provides a measure of transcriptional activity [2]. Sequential measurement of pol-II occupancy on genes released from transcriptional blockade, for example, in response to stimuli, reveal a wave of transcription moving through the body of the responding transcript.

A number of studies have attempted to determine the rate of transcription through modelling the dynamics of pol-II. Darzacq *et al.* fit a mechanistic model of pol-II transcription to nascent RNA data at a single locus and obtained a transcription speed of 4.3 kilobases per minute [4]. Wada *et al.* activated transcription of genes greater than 100 kbp in length and estimated the transcription speeds using a model that measures an intronic RNA signal through taking advantage of co-transcriptional splicing. They obtain an average transcription rate of 3.1 kbp min<sup>-1</sup> [5]. Singh and Padget (2009) reversibly inhibit transcription to determine the transcription rate of 10 genes, all of which were greater than 100 kbp which had an average transcription rate of 3.79 kbp min<sup>-1</sup> [6]. The data used in these studies have good temporal resolution (e.g. samples every 7.5 min in [5]) and reliably allows fitting of mathematical models or the direct measurement of transcription speed, however, only for a limited set of long genes. In contrast, high throughput data sets, such as ChIP-Seq, can be used to uncover transcription dynamics genome-wide but typically have much lower temporal resolution, motivating the development of alternative modelling approaches that report genome-wide transcription rates.

One way around the low temporal resolution of typical high-throughput time course data is to employ a non-parametric model of the biological signals of interest. In many cases we expect these signals to vary continuously and smoothly in time, when averaged over a cell population, and a Gaussian process model provides a convenient non-parametric model in such cases [7]. Gaussian processes have recently found applications in a range of biological system models [8–11].

Here we present a Gaussian process model of transcription dynamics which can be fitted to genome-wide pol-II occupancy data measured using ChIP-Seq. The model describes the movement of pol-II through the gene body and combines a flexible model of promoter-proximal pol-II activity with a reliable estimate of transcription speed. By identifying genes which fit the model well, we provide a useful method to identify actively transcribed genes. The model does not assume a constant transcription speed and can therefore identify variable rates of transcription, for example due to transcriptional pausing. Model parameters are determined using either maximum likelihood (ML) estimation or via Bayesian inference using Markov chain Monte Carlo (MCMC) sampling. The Bayesian approach provides confidence intervals for parameter estimates and can incorporate priors that capture domain knowledge, e.g. the expected range of transcription speeds, based on previous experiments.

We fit our model to a pol-II ChIP-Seq time course dataset from MCF7 breast cancer cells stimulated with estradiol. The model is used to identify the set of transcriptionally engaged genes and estimate their mean transcription rate and transcriptional activity near the promoter. By clustering promoter activity profiles, potential co-regulated groups of genes are identified, particularly those that respond rapidly to estrogen signalling. Subsequent characterisation of transcription factor (TF) binding sites in proximity to the promoters of genes within clusters provides a means of classifying groups of promoters that are responsive to the binding of specific combinations of TFs. Additionally, publically available ChIP-Seq datasets of TF profiles from the same system were used to identify cluster-specific patterns in TF-binding. The rates of transcription estimated by our model are consistent with the literature [4, 5] but with the advantage that our method allows the computation of transcription speeds genome-wide.

Our methodology has a number of advantages. We do not require data with high temporal resolution, making it feasible to model transcriptional dynamics genome-wide using ChIP-Seq or GRO-Seq time course data. We infer transcription rates for all genes in an unbiased manner and by using Bayesian parameter estimation we are able to associate our transcription rate estimates with confidence intervals. Our model is non-parametric and therefore does not make very strong assumptions about the temporal changes in transcriptional activity. Fitting the model genome-wide allows us to identify and filter out transcripts where pol-II does not travel down the gene body. This provides a principled method to identify responsive genes, in particular, early acting estrogen responsive genes in the specific application considered here. Since our model does not enforce a uniform transcription speed over the entire gene body, we can take into account phenomena such as pol-II pausing which would result in a non-uniform transcription speed. We also use this model to infer the promoter activity of transcriptionally engaged genes, to identify co-regulated gene modules downstream of estrogen signalling.

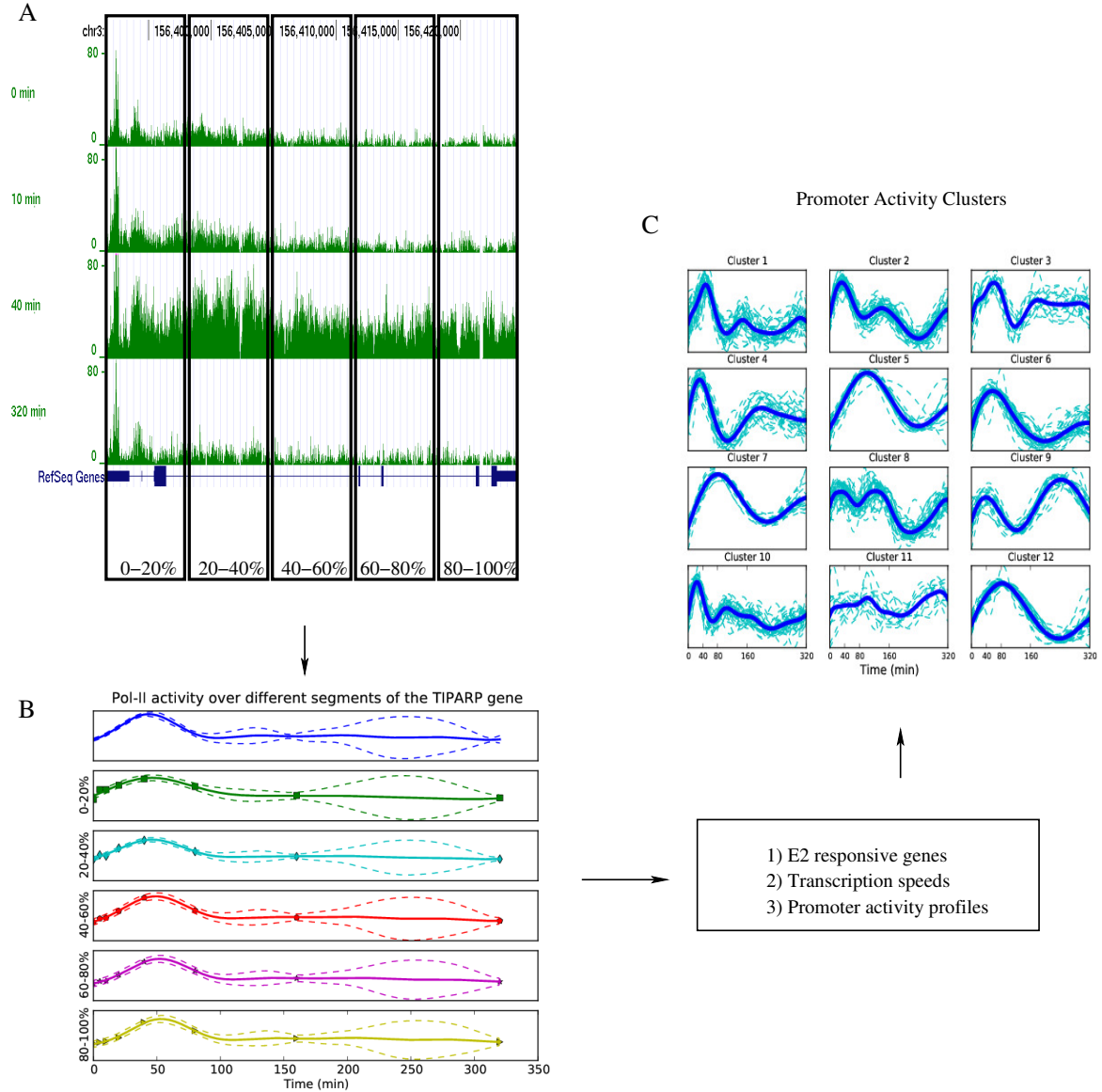
## Methods

Visualizing pol-II ChIP-seq reads mapped to transcriptional units at multiple time points following the addition of estradiol to MCF7 cells reveals the motion of pol-II through the gene body of estradiol responsive genes (see Figure 1). Computing the average pol-II occupancy over successive gene segments describes the motion of the transcription wave. Thereafter, fitting a model capable of smoothly interpolating between observed time points and by determining the time taken for pol-II to move from one gene segment to the next determines if pol-II is transcriptionally engaged on a given transcript and the speed at which it is moving through this transcriptional unit. We employ a Gaussian process formulation which takes into account the relationship between the pol-II signal at different regions of the gene and across time. Model parameters are determined using Maximum Likelihood (ML) or Bayesian inference (MCMC) to determine genes of interest and moreover, in the case of MCMC, determine confidence intervals for our parameter estimates.

### Gaussian Processes

A Gaussian process (GP) is a distribution over the space of functions. This distribution is completely specified by a mean function  $m(t)$  and a covariance function  $k(t, t')$ . A function  $f(t)$  is said to be drawn from a Gaussian process  $\mathcal{GP}(m(t), k(t, t'))$  if  $f(t)$  at any finite collection of points has a multivariate Gaussian distribution.

GPs are a powerful framework for non-parametric regression [7]. If a function is assumed to be drawn from a GP with known mean and covariance function, we can estimate the function value and associated uncertainty at unobserved locations given noise-corrupted observations. GPs have recently been applied in modelling biological systems, e.g. modelling protein concentrations as latent variables in differential equation models of transcriptional regulation [8, 9] and modelling spatial gene expression [11]. Here we introduce a novel application of GPs to modelling the spatio-temporal dynamics of pol-II occupancy



**Figure 1.** Pol-II ChIP-seq data for the TIPARP gene shows a transcription wave moving down the gene. The transcription dynamics model captures this motion and allows us to estimate transcription speeds. In this case the gene is divided into 5 segments and we estimate the speed to be approximately 2 kilobases per minute. Figure A shows the raw ChIP-seq reads at different times between 0 and 320 min. The top panel of Figure B shows the inferred promoter activity profile. The next five panels show the inferred profiles for the five gene segments corresponding to 0 – 20%, ..., 80% – 100% of the gene. By clustering these promoter activity profiles as shown in Figure C, we are able to group genes into clusters that are likely to be co-regulated and in particular we identify the clusters that respond most rapidly to estrogen signalling.

during transcription. The model makes use of a convolution of coupled GP functions across different

segments of the transcribed region.

## Convolved Gaussian Processes

Convolved GPs allow the modelling of correlations between multiple coupled data sources. In our case these data sources are the pol-II occupancy over time collected at different locations along the transcribed region of a gene. Modelling the data as a convolved process borrows information from these different data sources in estimating the model parameters and inferring the underlying signal in the data. The main idea is related to linear systems theory where the output  $y(t)$  of a linear time-invariant system whose impulse response is  $h(t)$  is given by the convolution of the input  $x(t)$  and  $h(t)$ , that is  $y(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau$ . If different sets of observations are believed to be related, they can be modeled as the outputs of different linear systems in response to a single input. If this input is modeled as a GP, then it will form a joint GP together with all the outputs and data from one output stream will be useful in inferring the rest [12–14].

## Model Description

In order to capture the movement of the transcription wave through transcriptional units, we divide each gene into  $I$  segments and compute time series of pol-II occupancy for each of the segments. Due to the low temporal resolution characteristic of high-throughput datasets, the time series between measurements must be inferred. To this end, we model the pol-II occupancy  $y_i(t)$  in each segment  $i \in \{1, \dots, I\}$  as the convolution of a latent process  $f(t)$  which is shared by all segments and a (possibly delayed) smoothing kernel  $k_i(\tau - D_i)$  corrupted by an independent white Gaussian noise process  $\epsilon_i(t)$  with zero mean and variance  $\sigma_i^2$  [13, 14]. That is

$$y_i(t) = \alpha_i \int_{-\infty}^{\infty} f(t - \tau)k_i(\tau - D_i)d\tau + \epsilon_i(t), \quad (1)$$

where  $\alpha_i$  is a scale factor and  $D_i$  is the delay of each segment. The latent process  $f(t)$  is modeled as a random function drawn from a GP with zero mean and a squared exponential covariance function (defined in Equation (4) below). The smoothing kernel is assumed to be Gaussian, that is

$$k_i(\tau) = \frac{1}{\sqrt{2\pi}\ell_i} \exp\left(-\frac{\tau^2}{2\ell_i^2}\right). \quad (2)$$

The estimated delay  $D_i$  of each smoothing kernel models the amount of time it takes the ‘transcription wave’ to reach the corresponding gene segment. This is used to estimate the transcription speed. Biologically the latent function can be thought of as modeling activity at the promoter while the smoothing kernel accounts for ‘diffusion’ of the transcription wave. This diffusion phenomenon is observed when time series of pol-II occupancy over different sections of a gene are plotted, with the transcription wave seen to spread out (see Figure 4). This phenomenon may be due to an initially synchronized cell population becoming less synchronized over time, resulting in broadening of the pol-II occupancy distribution over time.

Using equation (1), we can compute the covariance between the pol-II occupancy at various segments of the gene. We have

$$\text{cov}[y_i(t), y_j(t')] = \alpha_i \alpha_j \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k_f(t - \tau, t' - \tau')k_i(\tau - D_i)k_j(\tau' - D_j)d\tau d\tau' + \sigma_i^2 \delta_{ij} \delta_{tt'} \quad (3)$$

where

$$k_f(t, t') = \sigma_f^2 \exp\left(-\frac{(t - t')^2}{2\ell_f^2}\right). \quad (4)$$

Equation (3) can be evaluated in closed form using the fact that the product of two Gaussians yields an un-normalized Gaussian [7]. Exploiting this fact we get

$$\text{cov}[y_i(t), y_j(t')] = \alpha_i \alpha_j \frac{\sigma_f^2 \ell_f}{\sqrt{\ell_f^2 + \ell_i^2 + \ell_j^2}} \exp\left(-\frac{(t' - t + D_i - D_j)^2}{2(\ell_f^2 + \ell_i^2 + \ell_j^2)}\right) + \sigma_i^2 \delta_{ij} \delta_{tt'}. \quad (5)$$

Similarly,

$$\text{cov}[f(t), y_i(t')] = \alpha_i \frac{\sigma_f^2 \ell_f}{\sqrt{\ell_f^2 + \ell_i^2}} \exp\left(-\frac{(t' - t - D_i)^2}{2(\ell_f^2 + \ell_i^2)}\right). \quad (6)$$

## Inference

Let  $\mathbf{y}_i = [y_{i1}, \dots, y_{iN}]^\top$  be a vector of observations of pol-II occupancy over the  $i$ th gene segment and let  $\mathbf{Y} = [\mathbf{y}_1^\top, \dots, \mathbf{y}_I^\top]^\top$  be a vector formed by concatenating all the observations for a single gene.  $N$  is the number of observation time points and  $I$  is the number of gene segments so for a single gene  $\mathbf{Y}$  is a vector of length  $NI$ . We have

$$p(\mathbf{f}, \mathbf{Y} | \Theta) = \mathcal{N}([\mathbf{f}, \mathbf{Y}]; \mathbf{0}, \mathbf{K}), \quad (7)$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\mathbf{f}, \mathbf{f}} & \mathbf{K}_{\mathbf{f}, \mathbf{y}_1} & \dots & \mathbf{K}_{\mathbf{f}, \mathbf{y}_I} \\ \mathbf{K}_{\mathbf{y}_1, \mathbf{f}} & \mathbf{K}_{\mathbf{y}_1, \mathbf{y}_1} & \dots & \mathbf{K}_{\mathbf{y}_1, \mathbf{y}_I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{\mathbf{y}_I, \mathbf{f}} & \dots & \dots & \mathbf{K}_{\mathbf{y}_I, \mathbf{y}_I} \end{bmatrix} \quad (8)$$

and  $\Theta = \{\sigma_f, \ell_f, \{\alpha_i, D_i, \ell_i, \sigma_i\}_{i=1}^I\}$  are the parameters of our model which will be fitted on a gene by gene basis. The elements of  $\mathbf{K}$  are computed using equations (4), (5), and (6). By marginalizing over the latent function  $\mathbf{f}$ , we obtain the marginal likelihood  $p(\mathbf{Y} | \Theta)$ . Maximum likelihood estimates of the parameters  $\Theta$  are readily obtained by maximizing the log marginal likelihood using gradient-based optimisation.

For a fully Bayesian approach, we take advantage of the fact that the parameters are positive and bounded. We transform the parameters using a logit transform and work with unconstrained variables. We place a Gaussian prior over the parameters in the transformed domain and draw samples from the posterior using the Hamiltonian Monte Carlo (HMC) algorithm [15] (A more detailed description of the priors is included in the supplementary material).

## Results

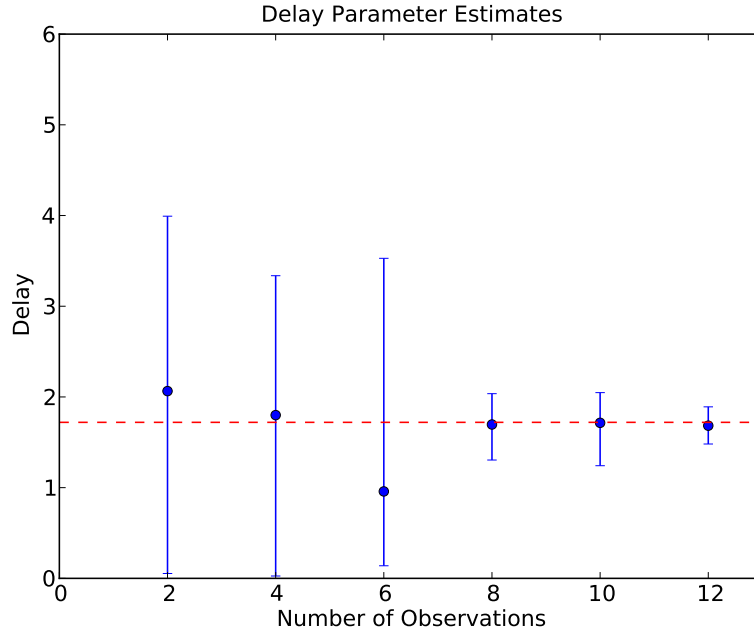
### Synthetic Data

We first apply our methodology to synthetic data generated according to the model. To generate the synthetic data, we use a latent function  $f(t)$  given by

$$f(t) = \sum_{i=1}^2 \beta_i \exp\left(-\frac{(t - \mu_i)^2}{2\sigma^2}\right)$$

with  $\beta_1 = \beta_2 = 1.5$ ,  $\mu_1 = 4$ ,  $\mu_2 = 5$  and  $\sigma = 1$ . We can analytically compute the integral in equation (1) to determine  $y_i(t)$ . We fixed the number of gene segments ( $I$ ) at two and generated random values of the parameters  $\{\alpha_i, D_i, \ell_i, \sigma_i\}_{i=1}^2$  with  $\alpha_i \in [0.5, 1]$ ,  $D_i \in [1, 2]$  and  $\ell_i \in (0, 1]$ . The values of  $\sigma_i$  were fixed at 0.05,  $D_1 = 0$  and  $\sigma_f = 1$  to avoid identifiability problems which arise if we try to estimate both  $\sigma_f$  and  $\{\alpha_i\}_{i=1}^2$ .

To determine the effect of number of observations  $N$  on the quality of inference, we compute the 95% confidence interval of the delay parameter as a function of number of observations using MCMC samples. Figure 2 shows the 95% confidence interval as a function of number of observations which are equally spaced between 0 and 16. In this case, the true value of  $D_2$  is 1.72. As expected the 95% confidence interval decreases and the median approaches the true value as we increase the number of observations. Figures 3(a) and 3(b) show the posterior distribution of  $f(t)$  and  $y_i(t)$  obtained with four and eight observations respectively using MCMC samples of the parameters.

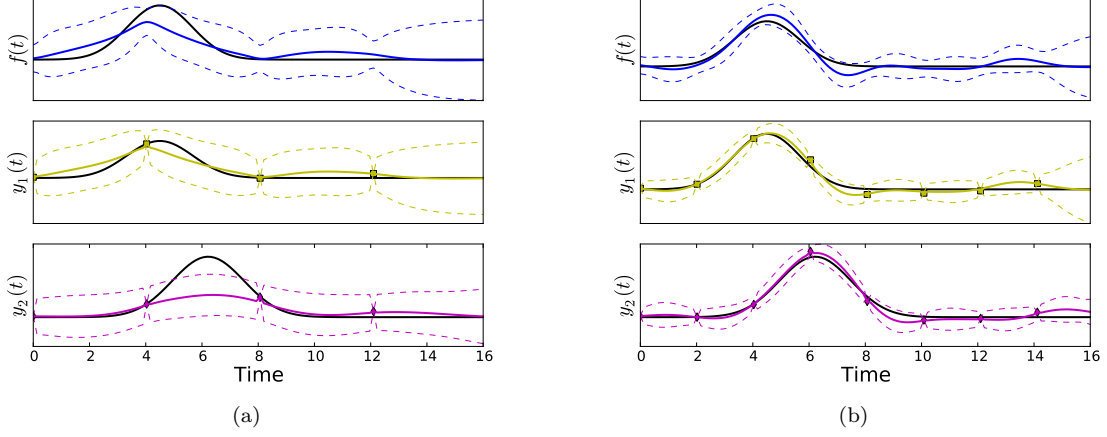


**Figure 2.** 95% confidence intervals for the delay parameter as a function of the number of observations. The true value of  $D_2$  is 1.72 (dashed red line).

## Estrogen Response Data

ChIP-seq was used to measure pol-II occupancy genome-wide when MCF-7 breast cancer cells are treated with estradiol (E2). Cells were put in estradiol free media for three days. This is defined media devoid of phenol red (which is estrogenic) containing 2% charcoal stripped foetal calf serum. The charcoal absorbs estradiol but not other essential serum components, such as growth factors. This results in basal levels of transcription from E2 dependent genes. The cells are then incubated with E2 containing media, which results in the stimulation of E2 responsive genes. The measurements were taken at logarithmically spaced time points 0, 5, 10, 20, ..., 320 minutes after E2 stimulation.

Raw reads were mapped onto the human genome reference sequence (NCBI\_build37) using the Genomatrix Mining Station (software version 3.2.1). The mapping software on the Mining Station is an index based mapper that uses a shortest unique subword index generated from the reference sequence to identify possible read positions. A subsequent alignment step is then used to get the highest-scoring match(es) according to the parameters used. We used a minimum alignment quality threshold of 92% for mapping and trimmed 2 basepairs from the ends of the reads to account for deterioration in read quality at the 3' end. The software generates separate output files for uniquely mapped reads and reads that have



**Figure 3.** Mean posterior distribution of  $f(t)$  and  $y_i(t)$  using synthetic data with four (a) and eight (b) observations using MCMC samples.

multiple matches with equal score. We only used the uniquely mapped reads. On average about 66% of all reads could be mapped uniquely. The data are available from the NCBI Gene Expression Omnibus under accession number GSE44800.

Time series of pol-II occupancy over various segments of genes were computed in reads per million (RPM) [16] using BEDtools [17,18]. The genes were divided into 200bp bins and the RPM computed for each bin. The occupancy in a particular gene segment was the mean RPM of the bins in that segment. Here, the gene is divided into five segments each representing 20% of the gene.

### Pol II Dynamics Genome-wide

For our initial experiment, we considered 3,064 genes which exhibit significant increase of pol-II occupancy between 0 and 40 minutes after E2 treatment. These genes were determined by counting the number of pol-II tags on the annotated genes in the RefSeq hg19 assembly at 0 and 40 minutes after E2 treatment and computing the  $\log_2$  ratio of these counts. We keep those genes where this quantity is greater than one standard deviation above the mean. For these 3,064 genes, we filtered out genes less than 1000bp in length and computed model fits using the ChIP-seq time series data for the remaining 2623 genes. The estimation of the parameters  $\{\sigma_f, \ell_f, \{\alpha_i, D_i, \ell_i, \sigma_i\}_{i=1}^5\}$  for a given gene was performed using maximum likelihood with  $D_1$  fixed at zero,  $\sigma_f = 1$  and the values  $\sigma_i$  tied to single value. Intuitively, one would expect the values of delay  $\{D_i\}_{i=1}^5$  to be non-decreasing. We therefore keep only those genes where this natural ordering is preserved for further analysis. We also discard genes with  $\hat{\ell}_f \leq 10$  and  $\hat{\ell}_f \geq 200$  since these are generally seen to be poor fits. Small values of  $\hat{\ell}_f$  arise when the data is best modelled as a noise process while large values model constant profiles which are not interesting in our analysis. This left us with 383 genes which we consider a conservative set of genes where there is evidence of engaged transcription. To rank these genes, we compared the log likelihood of the model fit to that obtained if we assume independence between the segments. This is equivalent to setting the off-diagonal blocks in equation (8) to the zero matrix.

The transcription speed is computed by assuming that the value of  $D_i$  is an indicator of how long it takes the ‘transcription wave’ to reach the corresponding gene segment. That is  $D_2$  is the amount of time it takes to transcribe 20% of the gene,  $D_3$  40% etc. To obtain confidence intervals on the speed estimates,



MCMC was performed to get samples of the parameters. Figures 4(a) - 4(f) show the inferred pol-II time profiles and histograms of the samples of the delay parameters for 3 of the top 10 genes found to fit the model well. To compute the transcription speed we plot the position along the gene in base pairs (bp) versus the delay in minutes and compute a linear regression through the origin. The slope of this line gives us the transcriptional speed. Each sample of the parameters provides a set of delay estimates from which we obtain a speed estimate. Figure 5(a) shows the linear regression plots using the delay samples for the *TIPARP* gene. Figure 5(b) shows the histogram of speed samples from which we can compute the confidence interval for the speed estimate. The 95% confidence interval is indicated in Figure 5(b) by the red triangle markers (cf. Table 1). Table 1 shows the transcription speeds for the top 10 genes computed using the samples of the delay parameters. Figure 6 shows a box plot of the transcription speeds computed using the samples of the delay parameters for these genes.

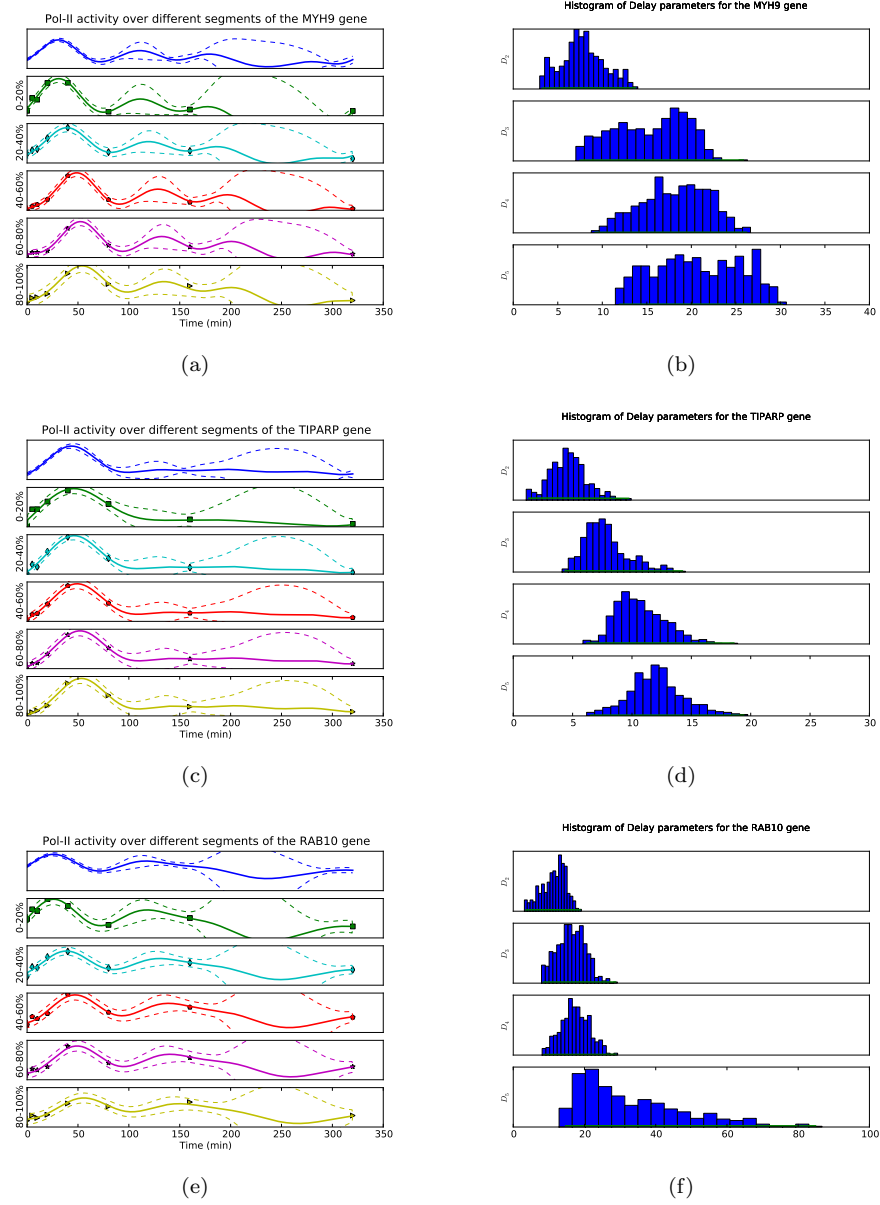
The advantage of fitting each of the delay parameters independently instead of enforcing a linear relationship is that it allows us to take into account phenomena such as pol-II pausing and provides a means to filter genes where the values of estimated delay are not naturally ordered. Visual inspection of the inferred time series of the top ranked genes is consistent with a ‘transcription wave’ traversing the gene. The transcription wave is especially evident in the longer genes *MYH9* and *RAB10*. This motivates a closer look at long genes. Table 2 shows the transcription speeds computed using the samples of the delay parameters for the 23 long genes found to fit the pol-II dynamics model well. Grouping these genes according to the magnitude of the median transcription speed allows us to compare our results to those presented previously. From Table 2 we see that 12 (52%) of these genes have transcription speeds between 2 and 4 kb per minute a range that includes the speeds reported in Wada *et al.* (3.1 kb per minute) [5] and Singh and Padget (2009) (3.79 kb per minute) [6].

Gene	Length (bp)	2.5%	50%	97.5%
TPM1	22196	1.6	2.4	4.1
WDR1	42611	1.0	1.6	3.5
TIPARP	32353	1.4	1.9	2.4
RHEB	53913	1.2	1.5	1.7
MYH9	106741	2.6	3.4	5.5
ACTN1	105244	0.6	2.8	4.2
PDLIM7	14208	1.7	3.5	6.4
ATP2A2	69866	3.6	6.8	10.2
RAB10	103595	1.4	2.6	4.4
AKAP1	36158	5.0	12.4	21.4

**Table 1.** Transcription speed in kilobases per minute for the top ten genes that fit the transcription model well. We use a Bayesian MCMC method for parameter estimation which provides the posterior distribution of the transcription speed. We show the 2.5%, 50% and 97.5% percentiles of the posterior distribution.

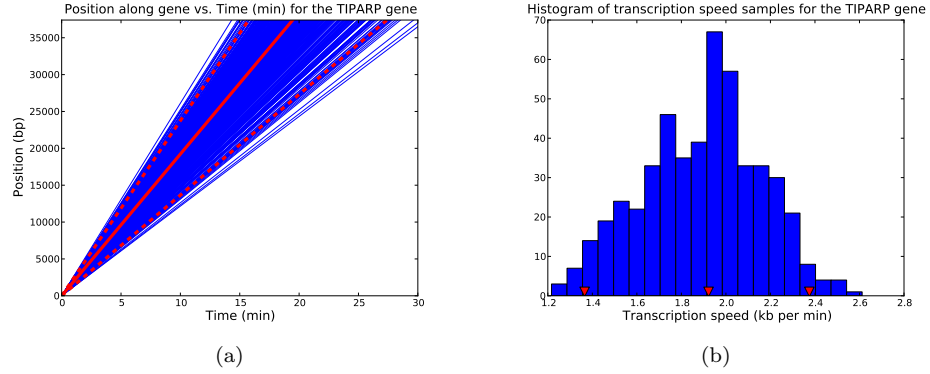
## Pathway analysis

To determine the biological significance of the 383 genes found to fit the pol-II dynamics model well, we used the Genomatix Pathway System (GePS) to look for enriched canonical pathways and gene ontology categories. Table 3 shows the significant canonical pathways ( $p$ -value  $< 0.05$ ) and the observed genes. It is interesting to note that the pair of genes *JAK1* and *JAK2* are responsible for a large number of the significant canonical pathways. These genes have previously been suggested as potential drug targets

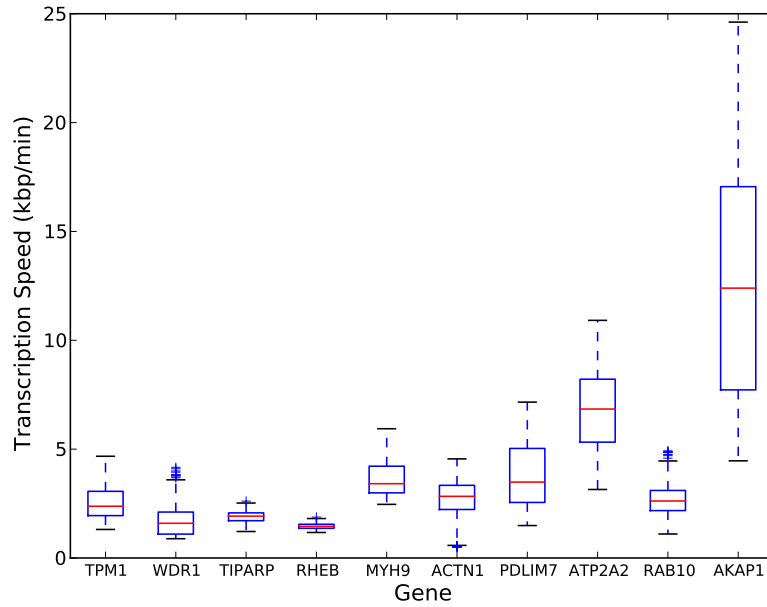


**Figure 4.** Inferred pol-II time profiles obtained for three of the top ten genes using ChIP-seq data. The top panel of each figure shows the inferred distribution of the latent function  $f(t)$ . The next five panels show the inferred profiles for the five gene segments corresponding to 0 – 20%, ..., 80% – 100% of the gene. The figures on the right are the delay histograms

in breast cancer (see for example [19]). The enrichment of the FOXA1 transcriptional network provides further confirmation that our model identifies biologically relevant genes. In recent work, Hurtado *et al.* [20] showed that FOXA1 influences the interaction of ER $\alpha$  and chromatin and therefore influences the response of breast cancer cells to E2. Genes in the FOXA1 canonical network found to fit the pol-II



**Figure 5.** Linear regression plots using the delay samples for the *TIPARP* gene (a) and the histogram of speed samples (b). The 95% confidence interval is indicated in (a) by the dashed red lines with the median represented by the solid red line. In (b) the 95% confidence interval is indicated by the red triangle markers (cf. Table 1).



**Figure 6.** Box plot of speed estimates for the top ten genes found to fit the transcription model well.

model well include *NRIP1* which is believed to be a direct E2 target that mediates the repression of ER $\alpha$  target genes later in the time course [21, 22]. Table 4 shows the top 20 significant gene ontology terms ( $p$ -value  $< 0.05$ ) for molecular function.

Gene	Length (bp)	2.5%	50%	97.5%
ACTN1	105244	0.6	2.8	4.2
ADCY1	148590	2.8	9.7	43.6
ARHGEF10L	158041	2.8	5.4	8.5
EPB41L1	120374	0.2	0.4	2.0
EPS15L1	110355	16.1	30.0	43.1
FARP1	102125	1.7	2.9	7.9
FLNB	163856	0.2	1.5	3.7
ITPK1	179005	0.3	2.9	6.8
JAK1	133282	0.6	2.2	4.2
JAK2	142939	0.6	2.4	5.3
KIAA0232	101441	0.9	2.3	4.0
KIF21A	150163	1.0	2.1	3.8
LARP1	104702	0.7	2.0	3.8
MYH9	106741	2.6	3.4	5.5
NCOR2	243050	6.5	10.9	20.5
NRIP1	103571	2.9	4.7	6.4
PKIB	116142	0.6	1.0	2.4
RAB10	103595	1.4	2.6	4.4
RAB31	154326	0.7	1.6	3.0
RASA3	150902	0.6	1.4	6.0
SHB	153316	0.5	3.1	5.0
WWC1	180244	1.9	3.6	5.6
ZNF644	106174	0.1	0.2	1.5

**Table 2.** Transcription speed in kilobases per minute for long genes between 100 and 300 kilobases long

### Clustering of Promoter Activity Profiles

The inferred latent functions for each gene model the pol-II activity adjacent to the promoter. Clustering these profiles and examining the average profiles of each cluster allows us to visualise the general trends and also classify genes according to the immediacy and nature of the response. This provides an alternative to clustering mRNA profiles. The production of mRNA may be delayed relative to the actual activation of transcription at the promoter causing genes which are actually triggered at the same time to show different rates of mRNA production. This could be due to their different lengths or due to the influence of processes such as pol-II pausing.

To classify the profiles we sample the mean of the latent function ( $f(t)$  in equation 1) and use PUMA-CLUST [23] to cluster the genes. PUMA-CLUST has the advantage of taking into account the uncertainty of the latent function when clustering the profiles. This uncertainty is computed from the covariance function inferred for  $f(t)$ .

The 383 genes found to fit the model well were grouped into 12 clusters. The number of clusters was determined using the Bayesian Information Criterion (see Figure 7) and the clusters are shown in Figure 8. To determine the speed of the response in each cluster, we compute the peak time of the mean profile for each cluster (see Table 5). Table 6 shows the significant canonical pathways ( $p$ -value  $< 0.01$ ) and the observed genes in each of the 12 clusters. The clustering of the pair of genes *JAK1* and *JAK2* in cluster 10, which has a prominent early peak, suggests that the response of both genes to E2 is rapid and coordinated. Since these genes are known to act together in several biological pathways, their appearance in the same cluster suggests that the clustering is likely to reveal other biologically significant relationships. One may also speculate that directly inhibiting the function of other early responding

Canonical pathway	Genes
IL-6 signaling pathway(JAK1 JAK2 STAT3)	JAK1, JAK2
IFN gamma signaling pathway	JAK1, JAK2
Proteasome complex	PSME1, PSMA4, PSMB5, PSMA2
IL-3 signaling pathway(JAK1 JAK2 STAT5)	JAK1, JAK2
Stat3 signaling pathway	JAK1, JAK2
FOXA1 transcription factor network	AP1B1, NDUFV3, NRIP1, SHH
PDGFR-alpha signaling pathway	JAK1, PDGFB, SHB
Hypoxia and p53 in the cardiovascular system	FHL2, HIF1A, GADD45A
LIF signaling pathway	JAK1, JAK2
IL-5 signaling pathway	JAK1, JAK2
p53 signaling pathway	TIMP3, GADD45A
IL-10 anti-inflammatory signaling pathway	JAK1, BLVRB
AndrogenReceptor	SPDEF, FHL2, STUB1 NCOR2, NRIP1
Integrin signaling pathway	CSK, ACTN1, NOLC1
Erythropoietin mediated neuroprotection through NF-KB	HIF1A, JAK2
PDGFR-beta signaling pathway	ACTR2, HCK, CSK, PDGFB, CTTN, JAK2
Mechanisms of transcriptional repression by dna methylation	RBBP7, MBD1
Hypoxia-inducible factor in the cardiovascular system	HIF1A, LDHA

**Table 3.** Significant canonical pathways.

genes would compromise the action of estradiol.

A closer look at the inferred pol-II promoter profiles of some examples in cluster 10, the earliest peaking cluster, and the corresponding inferred pol-II profiles over the last 20% of the genes reveals the possible influence of gene length on mRNA production and how clustering the inferred promoter profiles can account for this influence and uncover potential co-regulation. Figure 9 shows the inferred promoter profiles and the inferred pol-II profiles over the last 20% for three genes *CLN8*, *BRI3BP* and *JAK2* in cluster 10. Figure 10 shows the corresponding raw ChIP-seq reads. The lengths of the genes to the nearest kilobase are 23, 32 and 143 kb respectively. We see that despite the last segment profiles peaking at different times, the promoter profiles peak at approximately the same time. The difference in peak time over the final segment of the gene is most likely due to the length of the genes and accounts for the amount of time the pol-II takes to move down the gene. Such differences would mask potential co-regulation if we attempted to cluster genes based on their mRNA profiles.

In Hah *et al.* [3] GRO-seq was used to measure pol-II occupancy genome-wide when MCF-7 cells are treated with estradiol (E2) at four time points (0, 10, 40 and 160 min after E2 treatment). In addition, steady state levels of mRNA for 54 genes were measured using RT-qPCR at five time points (0, 10, 40, 160 and 320 min after E2 treatment). These data show a delay of between 1-3hr between peaks in the pol-II occupancy at the 5' end of a gene and peaks in the mRNA steady state [3, Figure S4]. These data include the mRNA measurement for 20 genes whose corresponding GRO-seq data peak is at 40 minutes after E2 treatment. Six of these genes namely *CASP7*, *FHL2*, *GREB1*, *ITPK1*, *NRIP1*, *WWC1* are found to fit our pol-II model well with ChIP-seq data. Table 7 shows the peak time of the inferred

Molecular function
Structural constituent of ribosome
RNA binding
Methyl-CpG binding
Protein binding
Structural molecule activity
Nucleic acid binding
rRNA binding
Non-membrane spanning protein tyrosine kinase activity
Ribosomal small subunit binding
Pseudouridine synthase activity
S100 alpha binding
Growth hormone receptor binding
Isomerase activity
Glucocorticoid receptor binding
Translation factor activity, nucleic acid binding
NF-kappaB binding
Threonine-type peptidase activity
Threonine-type endopeptidase activity
Intramolecular transferase activity

**Table 4.** Significant gene ontology terms.

Cluster	Peak Time (min)
1	48
2	32
3	61
4	32
5	100
6	58
7	80
8	122
9	242
10	22
11	297
12	80

**Table 5.** Cluster peak time

promoter profile  $T_p$ , the peak time of the inferred pol-II profile over the last 20% of the gene  $T_{20}$ , the GRO-seq peak time as well as the mRNA peak time. For the GRO-seq and mRNA peak times we show the peak times from [3, Figure S4] which are limited to the finite set of sampling times. We see that all mRNA peaks occur after  $T_{20}$ . The large value of  $T_{20}$  for *WWC1* which is a long gene  $\sim 180$  kb in length corresponds to a late peak in mRNA at 320 minutes. This shows that the parameters obtained by our model are biologically plausible. Based solely on the GRO-seq data these genes were grouped together in [3] since they show a peak at 40min. However our modeling reveals a greater diversity in the nature of responses. In fact the six genes appear in three different early response promoter profile clusters (see Table 7).

Cluster	Canonical pathway	Genes
1 (37)	PDGFR-beta signaling pathway	PDGFB, ACTR2, HCK
2 (47)	-	-
3 (18)	-	-
4 (29)	Nuclear receptors coordinate the activities of chromatin remodeling complexes and coactivators to facilitate initiation of transcription in carcinoma cells	NCOR2, TAF5
5 (27)	-	-
6 (40)	-	-
7 (24)	Proteasome complex Antigen processing and presentation	PSMB5, PSME1 PSMB5
8 (47)	-	-
9 (26)	-	-
10 (38)	IFN gamma signaling pathway IL-6 signaling pathway IL-3 signaling pathway Stat3 signaling pathway LIF signaling pathway IL-5 signaling pathway PDGFR-alpha signaling pathway IL27-mediated signaling events Role of ErbB2 in signal transduction and oncology IL6-mediated signaling events JAK_STAT_MolecularVariation.2	JAK2, JAK1 JAK2, JAK1 JAK2, JAK1 JAK2, JAK1 JAK2, JAK1 JAK2, JAK1 SHB, JAK1 JAK2, JAK1 JAK2, JAK1 JAK2, JAK1 JAK2, JAK1
11 (13)	-	-
12 (37)	-	-

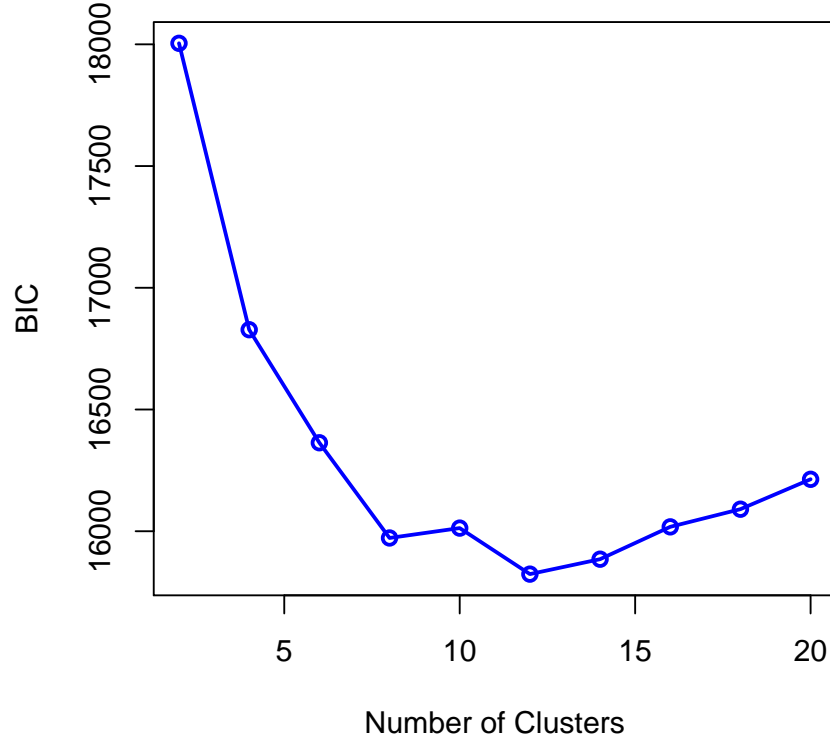
**Table 6.** Pathway analysis of clusters from inferred promoter activity profiles. The number in parentheses in column 1 is the cluster size.

Gene	Cluster	$T_p$	$T_{20}$	GRO-seq Peak	mRNA Peak
CASP7	1	36	47	40	160
FHL2	1	42	55	40	160
GREB1	2	30	46	40	320
ITPK1	2	36	64	40	160
NRIP1	10	22	40	40	160
WWC1	10	23	81	40	320

**Table 7.** The peak time of the inferred promoter profile  $T_p$ , the peak time of the inferred pol-II profile over the last 20% of the gene  $T_{20}$ , the GRO-seq peak time as well as the mRNA peak time (from [3, Figure S4]).

### Transcription factor binding

Tullai *et al.* [24] investigated genes that are co-regulated by shared transcription factor binding sites (TFBS). In particular, they found certain TFBS were enriched in the promoters of early response genes. We therefore investigated whether the promoters of genes in the different promoter profile clusters are enriched for different TFs. We use Pscan [25] to look for enriched TF motifs among those available in JASPAR [26]. The proximal promoter region -450 bp to +50 bp relative to the TSS of the genes in each

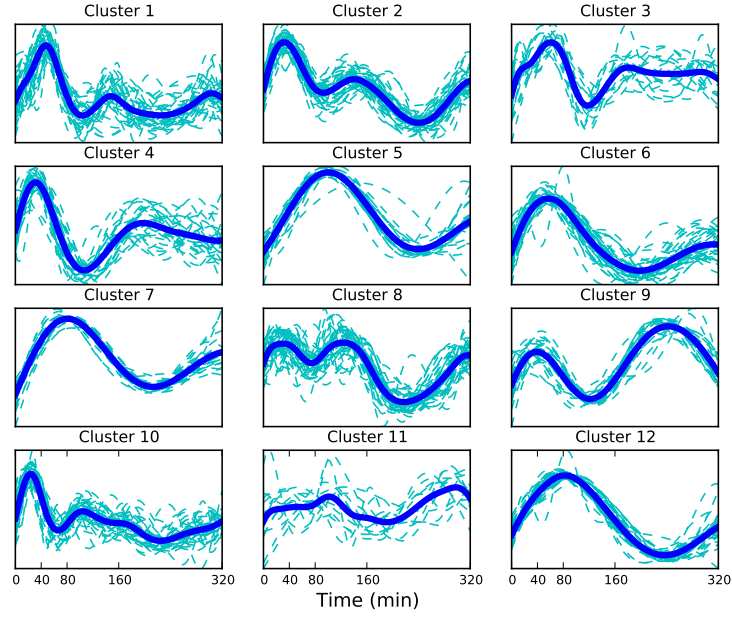


**Figure 7.** The BIC as a function of the number of clusters. The minimum BIC corresponds to 12 clusters.

cluster was analyzed. Table 8 shows significantly enriched TFBS in each cluster ( $p$ -value  $< 0.05$ ). Shown are TFs whose binding sites are over-represented in at least 5 clusters. The estrogen response element (ERE) is enriched in five clusters (1, 2, 5, 6 and 10), indicating that our modelling identifies estrogen responsive regions. The clusters containing an ERE have mean promoter activity profiles with distinct early peaks followed by decrease in activity which suggests transient activity. Additionally, clusters 1, 2 and 10 have relatively early peaks.

Next we investigated the EREs in the genes belonging to the 5 clusters enriched for the ERE motif. For each promoter sequence, the best sequence match to the ERE position frequency matrix (PFM) in JASPAR (MA0112.2) was determined. We keep those sequences with a matrix score greater than the mean score for matches found in the promoter sequences over the whole genome (For the ERE PFM this value is 0.73 when we consider the region -450 bp to +50 bp relative to the TSS). We used these sequences to determine the consensus ERE motif in this group of genes. To determine the consensus sequence, we report a single nucleotide for a given position if the nucleotide has a frequency greater than 50% and a frequency twice as large as the next nucleotide. We obtain a consensus sequence of 5'-GGnCACCTGnCC-3' (where n is any nucleotide) and an average matrix score of 0.77. The sequence





**Figure 8.** Clusters of promoter activity profiles. The mean profile in each cluster is shown by the bold line.

TF	Cluster											
	1	2	3	4	5	6	7	8	9	10	11	12
GABPA	✓	✓	✓	✓	✓	-	✓	✓	✓	-	✓	✓
Zfx	✓	✓	-	✓	✓	-	-	✓	✓	✓	✓	✓
Klf4	✓	✓	-	✓	-	-	✓	✓	✓	✓	-	✓
ELK1	✓	-	-	✓	✓	-	✓	✓	✓	-	✓	✓
HIF1A::ARNT	✓	✓	-	✓	✓	✓	✓	-	-	✓	✓	-
ELK4	✓	-	✓	✓	✓	-	✓	✓	✓	-	-	✓
SP1	✓	✓	-	✓	-	-	-	✓	✓	✓	-	✓
TFAP2A	✓	✓	-	✓	-	✓	-	✓	-	✓	-	-
Mycn	✓	-	-	✓	✓	-	-	✓	-	✓	✓	-
Myc	✓	-	-	✓	✓	-	-	✓	-	✓	✓	-
Pax5	✓	✓	-	✓	-	-	-	-	-	✓	-	✓
ER $\alpha$	✓	✓	-	-	✓	✓	-	-	-	✓	-	-
Arnt::Ahr	-	✓	-	✓	-	-	✓	✓	-	✓	-	-

**Table 8.** Significantly over-represented ( $p$ -value  $< 0.05$ ) transcription factor binding sites in the promoter profile clusters. We use Pscan to look for enriched TF motifs among those available in JASPAR. The proximal promoter region -450 bp to +50 bp relative to the TSS of the genes in each cluster was analyzed.

is visualised in Figure 11(a). The sequence of the ERE is known and given as 5'-GGTCAnnnTGACC-3' [27, 28]. The sequence corresponding to the PFM MA0112.2 is visualised in Figure 11(b). We see

that the ERE motif we obtain agrees well with the known motif (An analysis of the ERE motifs in the individual clusters is included in the supplementary material).

Determining the TFBS motifs enriched in each cluster provides a way to determine the influence of TFs on transcription. As a complementary approach, we also investigated the TF peaks in a 40 kbp region around the gene transcription start site for all genes in each cluster using ChIP-seq data for a number of TFs measured under similar experimental conditions (i.e. MCF-7 breast cancer cells treated with E2) in the cistrome database (<http://cistrome.org>). In earlier work on the estrogen interactome, Fullwood *et al.* [29] suggest that most long range interactions between TF binding sites and gene enhancers are limited to a range of about 20kb. We therefore investigate the region from -20kb to 20kb relative to the TSS (results for other regions around the TSS ranging from 1 to 100 kb are shown in the supplementary material). Table 9 shows the number of genes with TF binding peaks for each cluster for 7 TFs namely ER $\alpha$  [2], FoxA1 [30], c-Fos [31], c-Jun [31], c-MYC [32], SRC-3 [33], TRIM24 [34]. We found that the TFs RAD21 [35], CTCF [35] and STAG1 [35] are ubiquitously bound and not useful in uncovering cluster-specific TF binding. We investigate the statistical significance of the proportions of genes in each cluster with TF peaks in a 40kb neighborhood of the TSS by comparing the observed proportions to those we would expect in clusters of the same size drawn at random from the set of all genes. In Table 9 statistically significant ( $p$ -value  $< 0.05$ ) proportions are indicated in red (larger than expected) and green (lower than expected) with associated  $p$ -values in parentheses.

Interestingly, clusters 1, 2, 4, and 10, which show an early peak in the mean promoter profile, are all enriched for ER $\alpha$  and FOXA1. These clusters, with the exception of cluster 4, were also found to be enriched for the ER $\alpha$  motif near the promoter. The enrichment of both ER $\alpha$  and FOXA1 in these clusters is in line with conclusions drawn in Hurtado *et al.* [20] where it was suggested FOXA1 mediates ER $\alpha$  binding. We investigated the overlap of the binding sites for ER $\alpha$  and FOXA1 both in the 151 genes belonging to these clusters and genome-wide using the peaks obtained from [2] (ER $\alpha$ ) and [30] (FOXA1) and reported in the cistrome database. We investigated the 40kb region -20kbp to 20kbp relative to the TSS. Table 10 shows the number of ER $\alpha$  and FOXA1 peaks and the overlap (Two peaks are said to overlap if they have at least one base pair in common). We see that when we consider the rapid response genes in clusters 1, 2, 4, and 10 the percentage of overlap increases to 16% (35/220) whereas the overlap is 9% (956/11056) when we consider all genes. The significance associated with this elevated overlap is  $p=0.004$  given the null hypothesis of a random gene list of the same size (results for other regions around the TSS ranging from 1 to 100 kb are shown in the supplementary material). Taken together, the results in Tables 8, 9 and 10 identify genes that respond to E2, with clusters 1, 2, 4 and 10 most likely to contain the earliest E2 responsive genes.

## Discussion

In this work we have presented a methodology for modelling transcription dynamics and employed it to determine the transcriptional response of breast cancer cells to estradiol. To capture the movement of pol-II down the gene body, we model the observed pol-II occupancy time profiles over different gene segments as the delayed response of linear systems to the same input. The input is assumed to be drawn from a Gaussian process which models the pol-II activity adjacent to the gene promoter. Given observations from high-throughput data such as pol-II ChIP-Seq data, we are able to infer this input function and estimate the pol-II activity at the promoter. This allows us to differentiate transcriptionally engaged pol-II from pol-II paused at the promoter and yields good estimates of transcriptional activity.

In addition to estimating the transcriptional activity at the promoter, inferring the pol-II occupancy time profiles over different gene segments allows us to compute the transcription speed. We expect the delay parameters of different gene segments to be non-decreasing and this provides a natural way to determine genes that are being actively transcribed in response to E2.

Cluster	TFs						
	ER $\alpha$	FOXA1	c-FOS	c-JUN	MYC	SRC-3	TRIM24
1 (37)	<b>27</b> (0.001)	<b>14</b> (0.028)	<b>16</b> (0.001)	6	4	<b>25</b> (0.007)	27
2 (47)	<b>31</b> (0.003)	<b>19</b> (0.005)	<b>16</b> (0.022)	7	<b>7</b> (0.034)	<b>36</b> (0.001)	<b>38</b> (0.015)
3 (18)	11	5	7	<b>5</b> (0.029)	<b>6</b> (0.001)	11	12
4 (29)	<b>20</b> (0.007)	<b>11</b> (0.048)	9	<b>7</b> (0.023)	2	18	23
5 (27)	15	4	6	<b>8</b> (0.004)	<b>9</b> (0.001)	16	19
6 (40)	<b>27</b> (0.003)	8	12	7	4	<b>25</b> (0.027)	31
7 (24)	10	6	5	<b>6</b> (0.029)	3	13	19
8 (47)	<b>32</b> (0.001)	10	14	<b>14</b> (0.001)	<b>8</b> (0.011)	<b>31</b> (0.005)	<b>40</b> (0.002)
9 (26)	<b>18</b> (0.01)	7	<b>11</b> (0.01)	<b>11</b> (0.001)	3	12	<b>22</b> (0.025)
10 (38)	<b>30</b> (0.001)	<b>14</b> (0.036)	<b>15</b> (0.006)	2	1	<b>29</b> (0.001)	<b>32</b> (0.008)
11 (13)	5	2	<b>7</b> (0.008)	<b>4</b> (0.036)	2	7	<b>13</b> (0.004)
12 (37)	19	8	12	<b>11</b> (0.001)	4	<b>23</b> (0.037)	29

**Table 9.** Analysis of transcription factor binding in 40kbp regions of genes in gene clusters obtained from inferred promoter activity profiles. The number in parentheses in the first column is the cluster size. For each TF, we show the number of genes with peaks. Statistically significant proportions are indicated in red (larger than expected) with associated  $p$ -values in parentheses.

Genes	# of ER $\alpha$ peaks	# of FOXA1 peaks	ER $\alpha$ and FOXA1 overlap
Clusters 1, 2, 4, and 10 (151)	220 (112)	86 (44)	35 (0.004)
All genes ( $\sim 20,000$ )	11056	4626	956

**Table 10.** Overlap of ER $\alpha$  and FOXA1 binding in a 40 kb region around the TSS. The numbers in parentheses in the first column are the number of genes. In each TF peak column, we show the expected number of peaks in a set of random random genes of the same size in parentheses. In the overlap column the associated  $p$ -value is shown in parentheses.

Clustering the inferred promoter activity profiles allows us to investigate the nature of the response and group genes that are likely to be co-regulated. We found that the four clusters significantly enriched for both ER $\alpha$  and FOXA1 binding within 40kb according to public ChIP-Seq data were those that showed the earliest peak in pol-II activity at the promoter. ER $\alpha$  and FOXA1 ChIP peaks in the neighbourhood of these genes were also more likely to be overlapping than the average for ChIP-identified binding events of these TFs genome-wide. This observation provides some support for the previously proposed role of FOXA1 as a mediator of early transcriptional response in estrogen signalling.

As well as modelling transcriptional speed and transcriptional activity profiles, the proposed modelling approach may have other useful applications. For example, recent research has uncovered a link between transcription dynamics and alternative splicing [36]. It is believed that aberrant splicing can cause disease and a number of studies have tried to understand the mechanisms of alternative splicing [37]. The proposed model can potentially be used to identify transcriptional pausing events, and such results could be usefully combined with inference of splice variation from RNA-Seq datasets from the same system. Also, with the increasing availability of high-throughput sequencing data exploring multiple layered views of the transcription process and its regulation, the convolved modelling approach developed here has the potential to be usefully applied to more complex coupled spatio-temporal datasets.

## Acknowledgments

The work was funded by the European ERASysBio+ initiative project “Systems approach to gene regulation biology through nuclear receptors” (SYNERGY) by the BBSRC [BB/I004769/2 to CwM, MR and NDL], Academy of Finland [135311 to AH] and by the BMBF [grant award ERASysBio+ P#134 to GR]. We thank Nancy Bretschneider for running the mappings to generate the bed-files for this publication.

## Supplementary information

### Priors

The parameters  $\Theta = \{\sigma_f, \ell_f, \{\alpha_i, D_i, \ell_i, \sigma_i\}_{i=1}^I\}$  are positive and bounded. In the experiments we use the bounds shown in Table 11 with  $D_1$  fixed at zero,  $\sigma_f = 1$  and the values  $\sigma_i$  tied to single value. To determine the delay bounds, we assume that the value of  $D_i$  is an indicator of how long it takes the ‘transcription wave’ to reach the corresponding gene segment. That is  $D_2$  is the amount of time it takes to transcribe 20% of the gene,  $D_3$  40% etc. We obtain the length  $L$  of the gene from the hg19 annotation and use values of maximum and minimum expected speed ( $s_{min}$  and  $s_{max}$  respectively) to compute the delay bound. For example

$$D_2^{min} = \frac{0.2L}{s_{max}} \quad \text{and} \quad D_2^{max} = \frac{0.2L}{s_{min}}$$

We use  $s_{min}=50$  bp per min and  $s_{max} = 50\text{kb}$  per min. These large bounds allow unbiased estimation of transcription speed. (Recent work on individual cells suggests speeds as high as 50kb per minute are possible [38].)

We transform the parameters using a logit transform and work with unconstrained variables. For a parameter  $\theta \in \Theta$  with corresponding minimum and maximum bounds  $\theta_{min}$  and  $\theta_{max}$  respectively we compute the transformed variable  $\gamma$

$$\gamma = \log \left( \frac{\theta - \theta_{min}}{\theta_{max} - \theta} \right). \quad (9)$$

We place a Gaussian prior over the parameters in the transformed domain and draw samples from the posterior using the Hamiltonian Monte Carlo (HMC) algorithm [15]. We have

$$\gamma \sim \mathcal{N}(\gamma|0, \sigma_\gamma). \quad (10)$$

With  $\sigma_\gamma = 2$  we obtain an approximately uniform prior in the untransformed domain yielding an uninformative prior.

Parameter	Minimum	Maximum
$\ell_f$	5 min	320 min
$\alpha_i$	0	100
$D_i$	$\frac{0.2(i-1)L}{s_{max}}$ min	$\frac{0.2(i-1)L}{s_{min}}$ min
$\ell_f$	5 min	320 min
$\sigma_i$	0	100

**Table 11.** Parameter bounds.

To initialise the parameters for gradient optimisation, the length scales  $\ell_f$  and  $\ell_i$  are initialised at random from  $\{10, 20, 30, 40, 80\}$ ,  $\alpha_i$  and  $\sigma_i$  are drawn from  $\mathcal{U}[0, 1]$  with the value of  $\sigma_i$  multiplied by 100 to avoid local minima that would under-estimate the variance. The delays are initialised at random with the more realistic speed bounds  $s_{min}=500$  bp per min and  $s_{max} = 5\text{kb}$  per min when an ensemble of cells is considered. The parameters are then freely optimised with the bounds given in Table 11.

### Parameter gradients

To obtain ML estimates of the parameters we maximise the log marginal likelihood. To do this we require the gradients of the covariance function w.r.t the parameters. The gradients w.r.t  $\alpha_i$  and  $\sigma_i$  are straight

forward. Here we give the expressions for the gradients of  $\text{cov}[y_i(t), y_j(t')] = K_{yy}$  w.r.t  $\ell_f$ ,  $\ell_i$  and  $D_i$ . We have

$$\begin{aligned} \frac{\partial K_{yy}}{\partial \ell_f} &= \alpha_i \alpha_j \frac{\sigma_f^2(\ell_i^2 + \ell_j^2)}{(\ell_f^2 + \ell_i^2 + \ell_j^2)^{\frac{3}{2}}} \exp\left(-\frac{(t' - t + D_i - D_j)^2}{2(\ell_f^2 + \ell_i^2 + \ell_j^2)}\right) \\ &+ \alpha_i \alpha_j \frac{\sigma_f^2 \ell_f}{\sqrt{\ell_f^2 + \ell_i^2 + \ell_j^2}} \exp\left(-\frac{(t' - t + D_i - D_j)^2}{2(\ell_f^2 + \ell_i^2 + \ell_j^2)}\right) \frac{(t' - t + D_i - D_j)^2}{(\ell_f^2 + \ell_i^2 + \ell_j^2)^2} \end{aligned} \quad (11)$$

$$\begin{aligned} \frac{\partial K_{yy}}{\partial \ell_i} &= -\alpha_i \alpha_j \frac{\sigma_f^2 \ell_f \ell_i}{(\ell_f^2 + \ell_i^2 + \ell_j^2)^{\frac{3}{2}}} \exp\left(-\frac{(t' - t + D_i - D_j)^2}{2(\ell_f^2 + \ell_i^2 + \ell_j^2)}\right) \\ &+ \alpha_i \alpha_j \frac{\sigma_f^2 \ell_f}{\sqrt{\ell_f^2 + \ell_i^2 + \ell_j^2}} \exp\left(-\frac{(t' - t + D_i - D_j)^2}{2(\ell_f^2 + \ell_i^2 + \ell_j^2)}\right) \frac{\ell_i(t' - t + D_i - D_j)^2}{(\ell_f^2 + \ell_i^2 + \ell_j^2)^2} \end{aligned} \quad (12)$$

$$\frac{\partial K_{yy}}{\partial D_i} = -\alpha_i \alpha_j \frac{\sigma_f^2 \ell_f}{\sqrt{\ell_f^2 + \ell_i^2 + \ell_j^2}} \exp\left(-\frac{(t' - t + D_i - D_j)^2}{2(\ell_f^2 + \ell_i^2 + \ell_j^2)}\right) \frac{(t' - t + D_i - D_j)}{(\ell_f^2 + \ell_i^2 + \ell_j^2)} \quad (13)$$

To obtain gradient w.r.t the transformed parameters given by equation 9, we employ the chain rule.

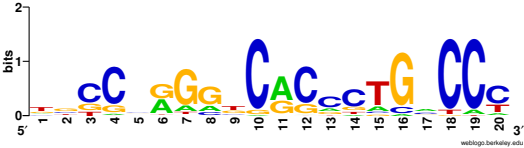
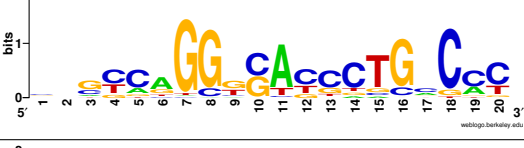

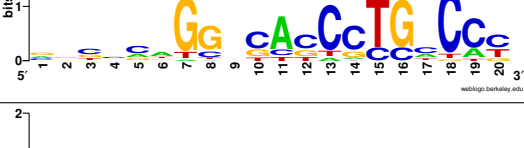
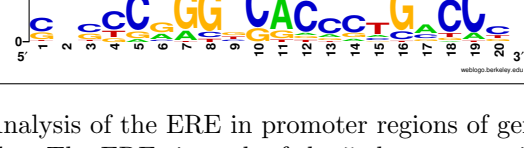
$$\begin{aligned} \frac{\partial K_{yy}}{\partial \gamma} &= \frac{\partial K_{yy}}{\partial \theta} \frac{\partial \theta}{\partial \gamma} \\ &= \frac{\partial K_{yy}}{\partial \theta} \frac{\exp(\gamma)(\theta_{max} - \theta_{min})}{(1 + \exp(\gamma))^2} \end{aligned} \quad (14)$$

## Motifs

Table 12 shows the EREs in each of the 5 clusters visualised using WebLogo. We also show the consensus sequence and the average matrix score. To determine the consensus sequence, we report a single nucleotide for a given position if the nucleotide has a frequency greater than 50% and a frequency twice as large as the next nucleotide. We see that there is some diversity in the motifs corresponding to different clusters but the consensus sequences agree with the known motif. Differences appear at at most 3 positions with the consensus sequence for cluster 10 differing at only two positions. We see that for the half site ‘TGACC’ the ‘A’ does not appear in the consensus sequence in all the clusters.

## Transcription factor binding

Tables 13 to 16 show the number of genes with TF binding peaks for regions around the TSS ranging from 1 to 100 kb for each cluster for 7 TFs namely ER $\alpha$  [2], FoxA1 [30], c-Fos [31], c-Jun [31], c-MYC [32], SRC-3 [33], TRIM24 [34]. In the tables, statistically significant ( $p$ -value  $< 0.05$ ) proportions are indicated in red (larger than expected) and green (lower than expected) with associated  $p$ -values in parentheses. These  $p$ -values are obtained empirically from 100,000 samples hence the cases where no extreme values are observed are reported as  $< 10e - 6$ .

Cluster	ERE Motif	Consensus sequence	Average Matrix Score
1		GnnCACCTGnCCC	0.772
2		GGnnACCCTGnCCn	0.77
5		GGnnAnCCTGnCCn	0.761
6		GGnnACCnTGnCCn	0.762
10		GGnCACCTGnCCn	0.765

**Table 12.** Analysis of the ERE in promoter regions of gene clusters obtained from inferred promoter activity profiles. The EREs in each of the 5 clusters are visualised using WebLogo (<http://weblogo.berkeley.edu/>). The consensus sequence is shown from position 7 which corresponds to the known ERE motif. The average matrix score is computed using the sequence matrix scores from Pscan.

We investigated the overlap of the binding sites for ER $\alpha$  and FOXA1 both in the 151 genes belonging to the rapid response genes in clusters 1, 2, 4, and 10 and genome-wide using the peaks obtained from [2] (ER $\alpha$ ) and [30] (FOXA1) and reported in the cistrome database. We investigated regions around the TSS ranging from 2 to 100 kb. Tables 17-20 show the number of ER $\alpha$  and FOXA1 peaks and the overlap. The statistical significance is determined by comparing the overlap in random gene lists of the same size

Cluster	TFs						
	ER $\alpha$	FOXA1	c-FOS	c-JUN	MYC	SRC-3	TRIM24
1 (37)	5	4	2	3	1	7	9
2 (47)	9 (8.2e-03)	3	2	2	4 (1.9e-02)	12 (1.3e-03)	10
3 (18)	3	2	2	1	3 (6.4e-03)	3	2
4 (29)	4	2	1	0 (< 10e-6)	0 (< 10e-6)	3	5
5 (27)	3	0 (< 10e-6)	0 (< 10e-6)	5 (9.7e-04)	4 (2.5e-03)	7 (1.1e-02)	5
6 (40)	5	3	3	0 (< 10e-6)	3	8 (3.4e-02)	2 (2.1e-02)
7 (24)	1	2	0 (< 10e-6)	3 (3.4e-02)	1	6 (2.3e-02)	7 (3.7e-02)
8 (47)	3	2	1	3	4 (1.9e-02)	6	14 (3.0e-03)
9 (26)	2	2	4 (2.5e-02)	5 (9.7e-04)	1	5	6
10 (38)	9 (1.9e-03)	2	1	0 (< 10e-6)	0 (< 10e-6)	3	9
11 (13)	0 (< 10e-6)	0 (< 10e-6)	3 (1.6e-02)	1	1	1	1
12 (37)	5	0 (< 10e-6)	2	5 (4.8e-03)	2	11 (5.8e-04)	7

**Table 13.** Analysis of transcription factor binding in 1kbp regions of genes in gene clusters obtained from inferred promoter activity profiles. The number in parentheses in the first column is the cluster size. For each TF, we show the number of genes with peaks. Statistically significant proportions are indicated in red (larger than expected) and green (lower than expected) with associated  $p$ -values in parentheses.

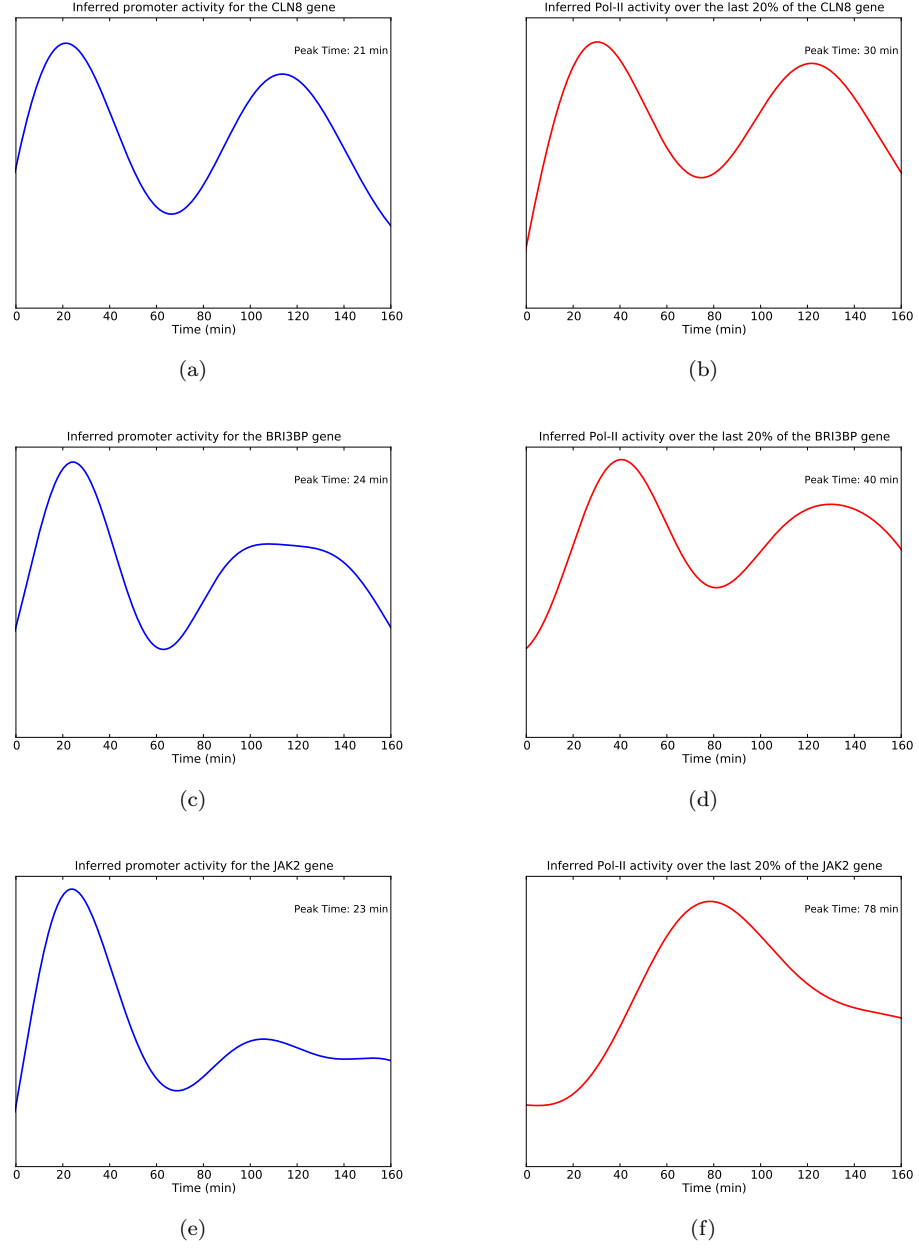
Cluster	TFs						
	ER $\alpha$	FOXA1	c-FOS	c-JUN	MYC	SRC-3	TRIM24
1 (37)	8 (3.2e-02)	4	3	3	1	8	10
2 (47)	10 (2.0e-02)	3	3	2	5 (7.1e-03)	14 (6.7e-04)	11
3 (18)	3	2	2	1	3 (1.0e-02)	3	3
4 (29)	4	2	1	0 (< 10e-6)	0 (< 10e-6)	3	9 (4.9e-02)
5 (27)	4	0 (< 10e-6)	1	5 (1.7e-03)	6 (1.0e-05)	8 (1.1e-02)	6
6 (40)	9 (1.8e-02)	5 (4.0e-02)	5	0 (< 10e-6)	3	11 (5.2e-03)	3 (2.3e-02)
7 (24)	2	3	0 (< 10e-6)	3 (4.5e-02)	1	7 (1.7e-02)	10 (4.4e-03)
8 (47)	5	2	1	4	4 (3.3e-02)	9	19 (1.9e-04)
9 (26)	3	2	6 (1.8e-03)	6 (1.4e-04)	1	7 (2.7e-02)	7
10 (38)	11 (1.1e-03)	3	2	0 (< 10e-6)	1	5	10
11 (13)	1	0 (< 10e-6)	3 (2.4e-02)	1	1	1	3
12 (37)	6	0 (< 10e-6)	2	5 (7.1e-03)	2	11 (2.6e-03)	8

**Table 14.** Analysis of transcription factor binding in 2kbp regions.

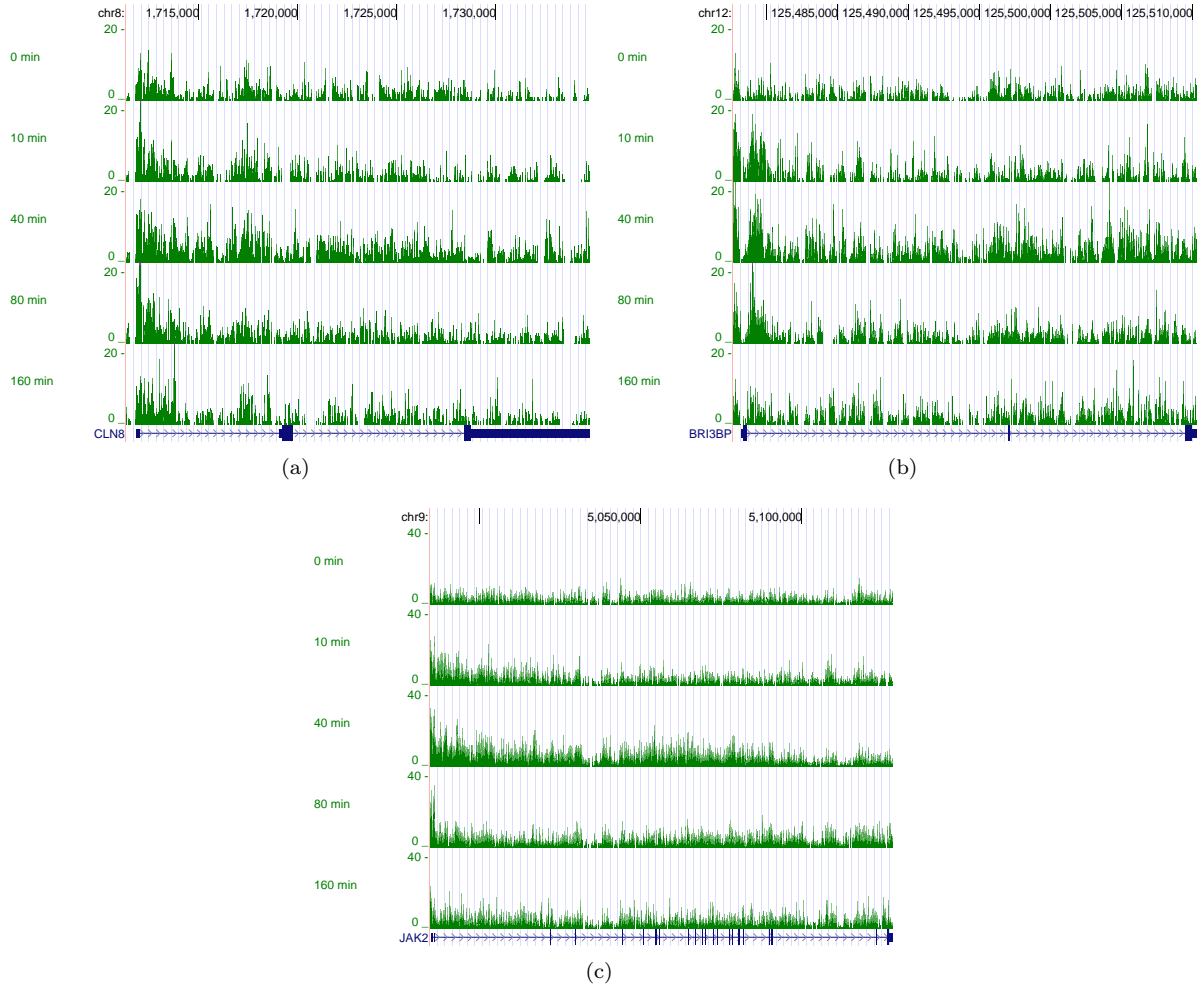
Cluster	TFs						
	ER $\alpha$	FOXA1	c-FOS	c-JUN	MYC	SRC-3	TRIM24
1 (37)	20 (1.9e-03)	9	8	4	1	18 (2.6e-02)	22
2 (47)	24 (2.3e-03)	13 (1.5e-02)	12 (1.5e-02)	6	7 (5.6e-03)	30 (1.0e-05)	28
3 (18)	4	4	4	2	5 (1.2e-03)	8	7
4 (29)	11	6	4	2	1	12	18
5 (27)	9	2	3	6 (9.2e-03)	8 (1.0e-05)	11	14
6 (40)	22 (9.3e-04)	8	6	4	3	18	24
7 (24)	7	4	2	4	2	13 (2.0e-02)	16 (4.8e-02)
8 (47)	21 (2.4e-02)	6	7	10 (1.2e-03)	7 (5.4e-03)	28 (6.0e-05)	34 (4.5e-04)
9 (26)	10	4	8 (1.5e-02)	9 (4.0e-05)	1	8	20 (1.9e-03)
10 (38)	26 (< 10e-6)	11 (1.7e-02)	9	0 (< 10e-6)	1	21 (2.6e-03)	24 (4.0e-02)
11 (13)	4	0 (< 10e-6)	5 (2.0e-02)	2	1	4	8
12 (37)	12	2 (2.2e-02)	7	10 (1.4e-04)	4	20 (4.8e-03)	23

**Table 15.** Analysis of transcription factor binding in 20kbp regions.

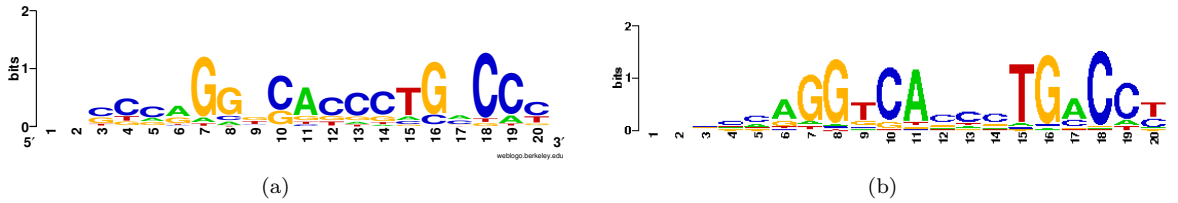




**Figure 9.** Inferred promoter profiles and pol-II activity over the final 20% of the gene for three genes in cluster 10.



**Figure 10.** ChIP-seq reads for three genes in cluster 10: *CLN8*, *BRI3BP* and *JAK2*.



**Figure 11.** Consensus sequence of regions matching the ERE motif in the promoter profile clusters enriched for the ERE motif (a) and the Estrogen Response Element (b).

Cluster	TFs						
	ER $\alpha$	FOXA1	c-FOS	c-JUN	MYC	SRC-3	TRIM24
1 (37)	29	20	<b>26</b> (4.0e-05)	<b>12</b> (2.1e-02)	4	<b>32</b> (8.0e-03)	<b>36</b> (2.2e-02)
2 (47)	<b>41</b> (3.0e-03)	<b>26</b> (4.1e-02)	23	11	<b>12</b> (7.0e-03)	<b>43</b> (1.3e-04)	43
3 (18)	<b>17</b> (1.1e-02)	7	10	6	<b>6</b> (1.4e-02)	14	16
4 (29)	<b>29</b> (1.0e-05)	<b>17</b> (4.9e-02)	15	<b>10</b> (2.1e-02)	5	<b>25</b> (2.1e-02)	28
5 (27)	21	8	11	<b>12</b> (1.1e-03)	<b>11</b> (1.9e-04)	19	24
6 (40)	<b>36</b> (1.5e-03)	15	19	11	6	<b>35</b> (3.6e-03)	38
7 (24)	15	11	8	<b>9</b> (1.7e-02)	5	18	22
8 (47)	<b>42</b> (9.5e-04)	20	22	<b>15</b> (1.3e-02)	9	<b>41</b> (2.0e-03)	<b>45</b> (2.5e-02)
9 (26)	<b>23</b> (1.9e-02)	15	<b>16</b> (8.0e-03)	<b>12</b> (7.2e-04)	5	<b>22</b> (4.6e-02)	24
10 (38)	<b>34</b> (2.6e-03)	<b>27</b> (1.8e-04)	<b>20</b> (3.1e-02)	5	4	<b>34</b> (1.9e-03)	36
11 (13)	9	4	8	4	2	10	13
12 (37)	<b>31</b> (3.1e-02)	<b>11</b> (4.7e-02)	<b>19</b> (4.5e-02)	<b>14</b> (2.8e-03)	5	28	35

**Table 16.** Analysis of transcription factor binding in 100kbp regions.

Genes	# of ER $\alpha$ peaks	# of FOXA1 peaks	ER $\alpha$ and FOXA1 overlap
Clusters 1, 2, 4, and 10 (151)	28 (12)	11 (6)	7 (0.042)
All genes ( $\sim 20,000$ )	1596	758	130

**Table 17.** Overlap of ER $\alpha$  and FOXA1 binding in a 1 kb region around the TSS. The numbers in parentheses in the first column are the number of genes. In each TF peak column, we show the expected number of peaks in a set of random genes of the same size in parentheses. In the overlap column the associated p-value is shown in parentheses.

Genes	# of ER $\alpha$ peaks	# of FOXA1 peaks	ER $\alpha$ and FOXA1 overlap
Clusters 1, 2, 4, and 10 (151)	36 (17)	13 (7)	8 (0.038)
All genes ( $\sim 20,000$ )	2220	929	177

**Table 18.** Overlap of ER $\alpha$  and FOXA1 binding in a 2 kb region around the TSS.

Genes	# of ER $\alpha$ peaks	# of FOXA1 peaks	ER $\alpha$ and FOXA1 overlap
Clusters 1, 2, 4, and 10 (151)	125 (63)	44 (26)	19 (0.045)
All genes ( $\sim 20,000$ )	7229	2991	626

**Table 19.** Overlap of ER $\alpha$  and FOXA1 binding in a 20 kb region around the TSS.

Genes	# of ER $\alpha$ peaks	# of FOXA1 peaks	ER $\alpha$ and FOXA1 overlap
Clusters 1, 2, 4, and 10 (151)	488 (254)	171 (100)	66 (0.006)
All genes ( $\sim 20,000$ )	17942	7927	1691

**Table 20.** Overlap of ER $\alpha$  and FOXA1 binding in a 100 kb region around the TSS.

## References

1. Hager GL, McNally JG, Misteli T (2009) Transcription dynamics. *Mol Cell* 35: 741-753.
2. Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FCGJ, et al. (2009) ChIP-Seq of ER $\alpha$  and RNA polymerase II defines genes differentially responding to ligands. *The EMBO Journal* 28: 1418-1428.
3. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, et al. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* 145: 622-634.
4. Darzacq X, Shav-Tal Y, de Turris V, Brody Y, Shenoy SM, et al. (2007) In vivo dynamics of RNA polymerase II transcription. *Nature structural & molecular biology* 14: 796-806.
5. Wada Y, Ohta Y, Xu M, Tsutsumi S, Minami T, et al. (2009) A wave of nascent transcription on activated human genes. *Proceedings of the National Academy of Sciences* 106: 18357-18361.
6. Singh J, Padgett RA (2009) Rates of in situ transcription and splicing in large human genes. *Nature Structural & Molecular Biology* 16: 1128-1133.
7. Rasmussen CE, Williams C (2006) *Gaussian Processes for Machine Learning*. MIT Press. URL <http://www.gaussianprocess.org/gpml/>.
8. Gao P, Honkela A, Rattray M, Lawrence ND (2008) Gaussian process modelling of latent chemical species: Applications to inferring transcription factor activities. *Bioinformatics* 24: i70-i75.
9. Honkela A, Girardot C, Gustafson EH, Liu YH, Furlong EEM, et al. (2010) Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences* 107: 7793-7798.
10. Kalaitzis AA, Lawrence ND (2011) A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics* 12.
11. Liu W, Niranjan M (2012) Gaussian process modelling for bicoid mrna regulation in spatio-temporal bicoid profile. *Bioinformatics* 28: 366-372.
12. Higdon D (2001) *Space and Space-Time Modeling Using Process Convolutions*. Technical report, Institute of Statistics and Decision Sciences, Duke University. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.5356>.
13. Boyle P, Frean M (2005) Dependent Gaussian processes. In: *In Advances in Neural Information Processing Systems* 17. MIT Press, pp. 217-224.
14. Alvarez M, Lawrence ND (2008) Sparse Convolved Gaussian Processes for Multi-output Regression. In: *NIPS*. pp. 57-64. URL [http://books.nips.cc/papers/files/nips21/NIPS2008\\_0170.pdf](http://books.nips.cc/papers/files/nips21/NIPS2008_0170.pdf).
15. Neal RM (2011) MCMC using Hamiltonian dynamics. In: S Brooks, A Gelman, G Jones and X-L Meng, editor, *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC.
16. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods* 6: S22-32.
17. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841-842.

18. Dale R, Pedersen B, Quinlan A (2011) Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics* .
19. The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490: 61–70.
20. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS (2011) FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics* 43: 27–33.
21. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics* 38: 1289–1297.
22. Jagannathan V, Robinson-Rechavi M (2011) Meta-analysis of estrogen response in MCF-7 distinguishes early target genes involved in signaling and cell proliferation from later target genes involved in cell cycle and DNA repair. *BMC Syst Biol* 5: 138.
23. Pearson R, Liu X, Sanguinetti G, Milo M, Lawrence N, et al. (2009) Puma: a Bioconductor package for propagating uncertainty in microarray analysis. *BMC Bioinformatics* 10: 211+.
24. Tullai JW, Schaffer ME, Mullenbrock S, Sholder G, Kasif S, et al. (2007) Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *J Biol Chem* 282: 23981-95.
25. Zambelli F, Pesole G, Pavesi G (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Research* 37: 247-252.
26. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an openaccess database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* 32: D91–D94.
27. Klinge CM (2001) Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Research* 29: 2905-2919.
28. Welboren WJ, Stunnenberg HG, Sweep FCGJ, Span PN (2007) Identifying estrogen receptor target genes. *Molecular oncology* 1: 138–143.
29. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor- $\alpha$ -bound human chromatin interactome. *Nature* 462: 58–64.
30. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, et al. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* 132: 958–970.
31. Joseph R, Orlov YL, Huss M, Sun W, Kong SLL, et al. (2010) Integrative model of genomic factors for determining binding site selection by estrogen receptor  $\alpha$ . *Molecular systems biology* 6: 456.
32. Hua S, Kittler R, White KP (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell* 137: 1259–1271.
33. Lanz RB, Bulynko Y, Malovannaya A, Labhart P, Wang L, et al. (2010) Global Characterization of Transcriptional Impact of the SRC-3 Coregulator. *Molecular Endocrinology* 24: 859-872.
34. Tsai WW, Wang Z, Yiu TT, Akdemir KC, Xia W, et al. (2010) TRIM24 links a non-canonical histone signature to breast cancer. *Nature* 468: 927–932.
35. Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, et al. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome research* 20: 578–588.

36. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, et al. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479: 74–79.
37. Tazi J, Bakkour N, Stamm S (2009) Alternative splicing and disease. *Biochimica et Biophysica Acta* 1792: 14–26.
38. Maiuri P, Knezevich A, De Marco A, Mazza D, Kula A, et al. (2011) Fast transcription rates of RNA polymerase II in human cells. *EMBO Rep* .